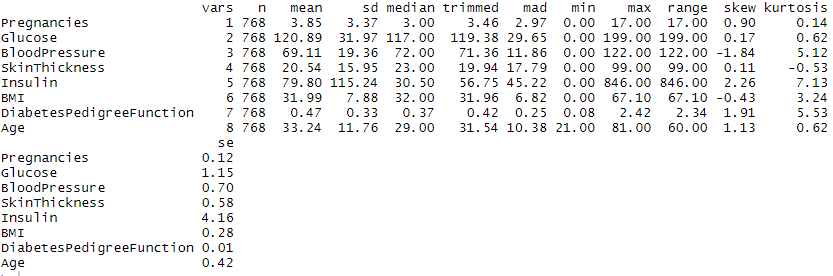
기초통계량 및 그래프

기초통계량과 그래프를 통해 개별 변수의 분포에 대한 기초 분석을 진행한다.



SVM model 선택 시, gamma값과 cost의 범위를 지정하였다.

Gamma는 2^c(-5:5)로 지정하였고 cost는 2^c(-4:4)로 지정하여 최적의 gamma와 cost의 조합을 찾았다.

SVM model에 대해 kernel trick을 주기 위해 polynomial, sigmoid, radial를 test 하였지만 polynomial은 시간이 너무 오래 걸려 결과가 중간에 나오지 않아서 사용이 부적합하다고 판단되었다.

결론적으로 SVM 모델이 당뇨병을 예측하는데 있어서 최적의 방법론으로 선택되었지만 직접 데이터를 가지고 프로그램을 돌려보았을 때 시간이 굉장히 오래 걸린다는 단점이 있었다.

현재 가지고 있는 데이터가 763개밖에 되지 않음에도 불구하고 시간이 오래 걸리는 것으로 보아 데이터 개수가 늘어난다면 SVM 모델의 사용이 어려울 것이다.

이때, accuracy나 error에 큰 차이가 없다면 lasso나 ridge penalty를 붙여서 logistic regression을 하는 것도 고려해볼만 하다.